PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-297766
(43)Date of publication of application : 18.11.1997

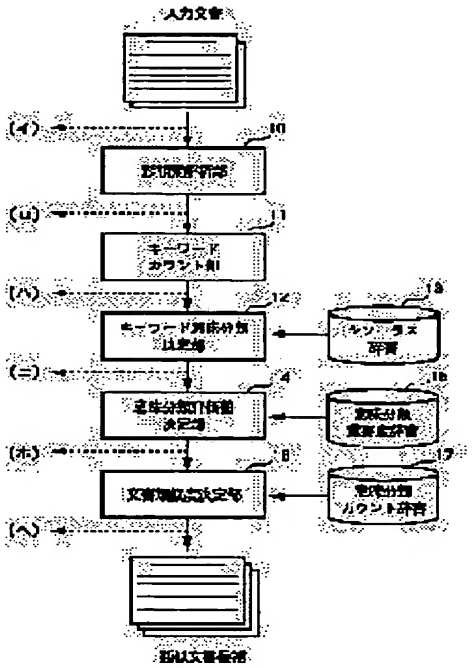
(51)Int.Cl. G06F 17/30

(21)Application number : 08-110870
(22)Date of filing : 01.05.1996

(71)Applicant : N T T DATA TSUSHIN KK
(72)Inventor : NAKAJIMA HIROYUKI
KITANI TSUYOSHI

(54) SIMILAR DOCUMENT RETRIEVAL DEVICE

(57)Abstract:
PROBLEM TO BE SOLVED: To provide the similar document retrieval device which can specify a document candidate similar to an input document among reference documents with high probability.
SOLUTION: The similar document retrieval device is constituted of a key work count part 11 which counts key words in the input document recognized by a morpheme analysis part 10, a key word semantic classification determination part 12 which sorts the key words included in the document according to semantic classifications, a semantic classification evaluated value determination part 14 which gives an evaluated value depending upon the importance, corresponding to a semantic classification and the number of key words belonging to each semantic classification, and a document similarity determination part 16 which gives similarity to each reference document according to the evaluated value.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号
特開平9－297766
(43)公開日　平成９年(1997)11月18日

(51)Int.Cl.[6]	識別記号	庁内整理番号	Ｆ Ｉ	技術表示箇所
Ｇ０６Ｆ　17/30			Ｇ０６Ｆ　15/403	３５０Ｃ
				３４０Ｂ

審査請求　未請求　請求項の数５　ＯＬ　（全７頁）

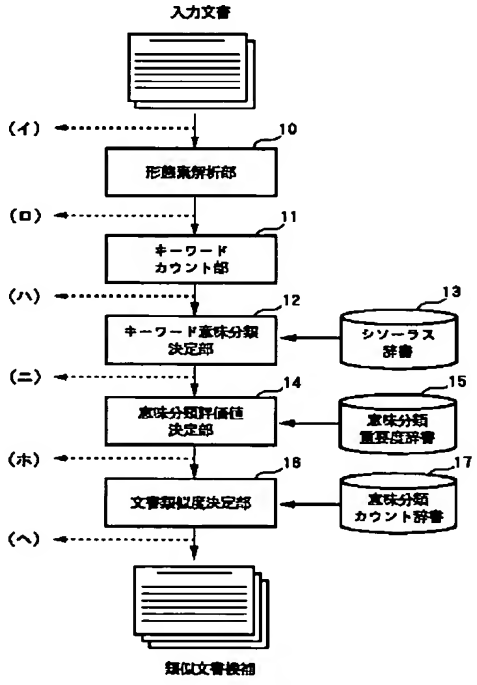
(21)出願番号　特願平8－110870

(22)出願日　平成８年(1996)５月１日

(71)出願人　000102728
エヌ・ティ・ティ・データ通信株式会社
東京都江東区豊洲三丁目３番３号
(72)発明者　中島　浩之
東京都江東区豊洲三丁目３番３号　エヌ・ティ・ティ・データ通信株式会社内
(72)発明者　木谷　強
東京都江東区豊洲三丁目３番３号　エヌ・ティ・ティ・データ通信株式会社内
(74)代理人　弁理士　鈴木　正剛

(54)【発明の名称】　類似文書検索装置

(57)【要約】
【課題】　複数の参照用文書から入力文書に類似する類似文書候補を高い確度で特定し得る類似文書検索装置を提供する。
【解決手段】　形態素解析部１０により認識された入力文書中のキーワードの個数を計数するキーワードカウント部１１、文書に含まれるキーワードを意味分類毎に仕訳するキーワード意味分類決定部１２、意味分類に応じた重要度と各意味分類に属するキーワードの個数に依存する評価値を付与する意味分類評価値決定部１４、及び評価値に基づいて各参照用文書毎に類似度を付与する文書類似度決定部１６とを含んで類似文書検索装置を構成する。



1

【特許請求の範囲】

【請求項１】　入力文書の所定領域を形態素解析して該入力文書に記述された語句を自動認識する文字認識手段と、

複数の参照用文書から前記認識された語句に関連する少なくとも一つの類似文書候補を特定する類似文書特定手段とを備え、

前記類似文書特定手段は、前記文字認識手段で認識された語句群をそれぞれ相異なる値に重み付けられた複数のグループに分類仕訳する第１手段、個々のグループに分類された語句の数に応じて各グループの重み評価値を演算する第２手段、及び前記複数の参照用文書に含まれるグループの各々に前記演算手段より算出された重み評価値を付与して各参照用文書を差別化する第３手段を含んで構成されていることを特徴とする類似文書検索装置。

【請求項２】　前記第１手段は、単一グループに属する語句名と当該グループ名、及び複数グループに属する可能性のある語句名と各グループ名がそれぞれ対応付けられた第１辞書と、前記文字認識手段で認識された語句が属するグループを前記第１辞書を照合して決定するグループ決定部とを有し、該グループ決定部は、複数グループに属する可能性のある語句については対応関係にある各単一グループに属している語句数に応じていずれか一のグループを決定することを特徴とする請求項１記載の類似文書検索装置。

【請求項３】　前記第２手段は、個々のグループに属する語句数の増加に伴い当該グループについての重み評価値を高くするように構成されていることを特徴とする請求項１または２記載の類似文書検索装置。

【請求項４】　前記第３手段は、複数の参照用文書の各々の文書識別コードと各参照用文書に含まれるグループ別の語句数とを対応付けて蓄積した第２辞書と、

第２辞書に蓄積されている各グループにそれぞれ前記算出された重み評価値を付与して文書識別コード毎の総合評価値を導出するとともに、この総合評価値の相対的大小に応じて参照用文書の前記入力文書への類似度を決定する類似度決定部と、を有することを特徴とする請求項１ないし３のいずれかの項記載の類似文書検索装置。

【請求項５】　前記グループは、語句の利用目的に応じた重み係数が付与された意味グループであることを特徴とする請求項１ないし４のいずれかの項記載の類似文書検索装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、例えば種々の情報検索システムや文書作成支援システム等に使用される文書検索技術に係り、特に複数の参照用文書から入力文書に含まれる語句に関連する文書候補を検索する類似文書検索装置に関する。

【０００２】

2

【従来の技術】類似文書検索装置は、予め蓄積されている複数の参照用文書から入力文書の文章内容により合致する、即ち類似する文書候補を検索する装置である。従来より、この種の類似文書検索装置では、入力文書の所定領域を形態素解析して文字、数字、記号、語、語句（以下、この明細書では語句と称して説明する）を認識するとともに、該入力文書に含まれる文章を特定の意味表現を有するいくつかのキーワードに分解した上で、例えば検索目的に応じた重要度のキーワードを含む度合いが高い少なくとも一つの文書候補を複数の参照用文書から選び出している。

【０００３】具体例を図７から図９を参照して説明する。図７は、従来のこの種の類似文書検索装置の構成例を示す図である。この類似文書検索装置は、形態素解析部７１、キーワードカウント部７２、キーワード評価値決定部７４、文書類似度決定部７６、及びキーワード重要度辞書７３のほか、キーワードカウント辞書７５などの辞書類を含んで構成される。

【０００４】いま、この類似文書検索装置に、図９（ト）に例示する文章「類似した文書を検索する装置を開発・・・・」が記述されている入力文書が入力されたとする。形態素解析部７１は、この文章の所定領域を形態素解析してキーワードを認識するとともに各キーワードにその品詞情報「類似（サ変名詞）」，「文書（名詞）」・・・付する。図９（チ）は形態素解析部７１の出力例を示すものである。

【０００５】キーワードカウント部７２は、形態素解析部７１の出力から上記文章に含まれる各々のキーワードの個数をカウントするとともに、キーワード別にその個数を出力する。図９（リ）は、キーワードカウント部７２の出力例を示すものである。キーワード評価値決定部７４は、キーワードカウント部７２からの出力（キーワードの個数）と、図８（ａ）に例示されるように予めキーワード毎に定められた重要度「類似：１００」，「検索：５０」・・・が記録されているキーワード重要度辞書７３とを参照しながら各キーワードに対する評価値を計算するとともに、この評価値を出力する。

【０００６】評価値は、各キーワードに対して与えられた重要度と入力文書に含まれるキーワードの個数の積で得られる数、あるいは、重要度と個数の対数の積で得られる数とする方法が考えられている。この方法は、ＴＦ／ＩＤＦ法と呼ばれているもので、Mc-Graw-Hill Publishing Company　から出版されているGerard Salton等による著書“Introduction to Modern Information Retrieval”の記載が参考になる。図９（ヌ）は、前者の手法を用いたキーワード評価値決定部７４の出力例を示すものである。

【０００７】文書類似度決定部７６は、キーワード評価値決定部７４からの出力と、図８（ｂ）に例示されるようにどのキーワードがどの参照用文書に何個含まれてい

るかが記録されたキーワードカウント辞書７５とを参照して、各参照用文書のそれぞれに対して類似度を付与する。また、類似度の高い順に類似文書候補とずる。この文書類似度決定部７６における参照用文書の類似度は、キーワードの重要度と各参照用文書中のキーワードの個数から上記ＴＦ／ＩＤＦ法などによって決定される。実際には、検索処理時間の短縮のために相対的に評価値の高いキーワードを適当に取り出して類似度の計算に使用する。図９（ル）は、この文書類似度決定部７６の出力例であり、各参照用文書に対応する文書識別コード、例えば文書番号を類似度の高い順に類似文書候補とした場合の例が示されている。

【０００８】

【発明が解決しようとする課題】上述の類似文書検索装置では、予め固定的に与えられたキーワード重要度辞書７３を参照しているため、複数の意味が派生するキーワードに対して入力文書中における本来の意味を考慮されることなく不当な重要度が与えられる可能性が高い。そのため、不当な重要度に基づいて評価値及び類似度の決定を行った場合に、確度の高い類似検索ができないといった問題があった。

【０００９】このことを簡単な例を挙げて説明する。上述の類似文書検索装置のキーワード重要度辞書７３において、例えばキーワードＡの重要度が”１０”、キーワードＢの重要度が”１００”、キーワードＣの重要度が”１０”で与えられており、キーワードＣは、それぞれ意味内容の異なるキーワードＡ及びキーワードＢの共通の短縮表現で使われるものとする。この場合に、入力文書に含まれるキーワードＢがキーワードＣに置き換えられていると、キーワード重要度辞書７３を使用する際に、その文書は、キーワードＢ”１００”に関してはそれと同義のキーワードＣの重要度”１０”をもって評価されて、不当に低いものとなる。そのため、キーワード文書類似度決定部７６における類似度の決定精度が悪くなる。

【００１０】逆に、キーワード重要度辞書７３においてキーワードＡが重要度”１０”であるのに対し、キーワードＣがキーワードＢと同様の重要度”１００”が与えられている場合、キーワードＡがその短縮表現であるキーワードＣに置き換えられている入力文書を扱うと、その入力文書中のキーワードＡ”１０”に関してはそれと同義のキーワードＣの重要度”１００”をもって評価されて、不当に高いものとなる。この場合も、文書類似度決定部７６における類似度の決定精度が悪くなる。

【００１１】本発明の課題は、入力文書に含まれる各キーワードに正当な重みを付与して文書間の類似度を正しく決定する類似文書検索装置を提供することにある。

【００１２】

【課題を解決するための手段】上記課題を解決するため、本発明は、入力文書の所定領域を形態素解析して該

入力文書に記述された語句を自動認識する文字認識手段と、複数の参照用文書から前記認識された語句に関連する少なくとも一つの類似文書候補を特定する類似文書特定手段とを備えた類似文書検索装置を提供する。この類似文書検索装置において、類似文書特定手段は、文字認識手段で認識された語句群をそれぞれ相異なる値に重み付けられた複数のグループ、例えば語句の利用目的に応じた重み係数が付された意味グループに分類仕訳する第１手段、個々のグループに分類された語句の数に応じて各グループの重み評価値を演算する第２手段、及び複数の参照用文書に含まれるグループの各々に前記演算手段より算出された重み評価値を付与して各参照用文書を差別化する第３手段を含んで構成することを特徴としている。

【００１３】本発明の類似文書検索装置における好ましい態様として、前記第１手段ないし第３手段を以下のように構成する。第１手段を、例えば単一グループに属する語句名と当該グループ名、及び複数グループに属する可能性のある語句名と各グループ名がそれぞれ対応付けられた第１辞書と、前記文字認識手段で認識された語句が属するグループを前記第１辞書を照合して決定するグループ決定部とを備えて構成し、該グループ決定部は、複数グループに属する可能性のある語句については対応関係にある各単一グループに属している語句数に応じていずれか一のグループを決定するようにする。第２手段を、個々のグループに属する語句数の増加に伴い当該グループについての重み評価値を高くするように構成する。第３手段を、複数の参照用文書の各々の文書識別コードと各参照用文書に含まれるグループ別の語句数とを対応付けて蓄積した第２辞書と、第２辞書に蓄積されている各グループにそれぞれ前記算出された重み評価値を付与して文書識別コード毎の総合評価値を導出するとともに、この総合評価値の相対的大小に応じて参照用文書の前記入力文書への類似度を決定する類似度決定部と、を備えて構成する。

【００１４】

【発明の実施の形態】以下、複数の参照用文書から入力文書に含まれる文章の意味表現に類似するものを索出する類似文書検索装置を例に挙げて本発明の実施形態を説明する。図１は、本発明の一実施形態に係る類似文書検索装置の構成図であり、語句としてキーワード、グループとしてキーワードの意味を表す意味分類を用い、さらに第１辞書として、類義語を各キーワードがそれぞれ属する意味分類と対応付けて蓄積したシソーラス辞書を用いる場合の例を示すものである。

【００１５】図１に示されるように、本実施形態の類似文書検索装置は、形態素解析部１０、キーワードカウント部１１、キーワード意味分類決定部１２、意味分類評価値決定部１４、文書類似度決定部１６のほか、キーワード意味分類決定部１２が参照するシソーラス辞書１３

と、意味分類評価値決定部１４が参照する意味分類重要度辞書１５と、文書類似度決定部１６が参照する意味分類カウント辞書１７とを含んで構成される。辞書１３～１５を除く各部１０，１１，１２，１４，１６は、例えばプログラムされたコンピュータ内に実現される機能モジュールである。この場合、辞書１３～１５及び各機能モジュールを各々独立した装置構成体ないしシステム構成体として存在させてもよく、あるいはコンピュータに諸機能を付与し得るようにするため、コンピュータが読み取り可能な態様で共通の情報記憶媒体に固定させるようにしてもよい。要は、コンピュータが稼働したときに、上記各機能モジュールが該コンピュータ内に形成されればよい。本実施形態では、従来例との相違を明確にするため、便宜上、各々独立した装置構成体として各機能モジュールが存在するものとして説明する。

【００１６】この類似文書検索装置の動作例を図２ないし図６を参照して説明する。いま、図３（イ）に例示する文章「マッキントッシュをマックと呼ぶことが流行している。・・・」が記述されている文書が入力されたとする。ここにマッキントッシュはアップル社のコンピュータの商標であり、「マック」はその愛称である。形態素解析部１０は、この入力文書の所定領域を形態素解析して文章中の語句を認識するとともに、文章からキーワード単位「マッキントッシュ」，「マック」・・・を抽出してそれぞれ品詞情報を付する。図３（ロ）は形態素解析部１０の出力例を示すものである。

【００１７】キーワードカウント部１１は、形態素解析部１０の出力に基づいて入力文書中の各キーワードの個数をカウントする。図３（ハ）は、キーワードカウント部１２の出力例であり、ここでは、入力文書中に「マック」が２個、「マッキントッシュ」が１０個・・・が含まれていたことが示されている。キーワード意味分類決定部１２は、シソーラス辞書１３を参照して、キーワードカウント部１１の出力結果を分類仕訳し、キーワード個数を意味分類毎に出力する。

【００１８】シソーラス辞書１３の内容例を図２（ａ）に示す。図示の例では、キーワード「マッキントッシュ」が意味分類「１」、キーワード「マクドナルド（マクドナルド社の商標：以下同じ）」が意味分類「２」、キーワード「マック」が意味表現「１，２」、キーワード「流行」が意味分類「３」に属している。なお、「マック」の意味分類「１，２」は、「マック」のキーワードが「マッキントッシュ」と「マクドナルド」の双方に属する可能性があることを意味している。上記シソーラス辞書１３を参照した場合の意味分類の特定と各意味分類毎のキーワード個数のカウント処理の概要を示したものが図４である。

【００１９】図４を参照すると、まず、「マッキントッシュ」や「マクドナルド」のように一つの意味分類のみに属するキーワードについて、その意味分類毎にキーワード個数を加算する。次に、「マック」のように複数の意味分類に属しているキーワードについては、最も合計個数の多い意味分類に属する一のキーワードを決定することで、それが属する意味分類を一つに絞る。例えば図４の例では、キーワード「マッキントッシュ」が１０個であるのに対し、キーワード「マクドナルド」は０個である。従って、「マック」のキーワードは、「マッキントッシュ」の意味分類「１」に属すると決定し、その個数”２”を意味分類「１」に加算する。その結果、キーワード意味分類決定部１２では、意味分類「１」に属するキーワードが１２個、意味分類「３」に属するキーワードが５個のように決定する。このキーワード意味分類決定部１２の出力例を図３（二）に示す。なお、キーワード意味分類決定部１２において、全てのキーワードを処理する代わりに、個数の多いキーワードを選択的に処理することも可能である。

【００２０】意味分類評価値決定部１４は、意味分類重要度辞書１５を参照して、上記キーワード意味分類決定部１２から出力された意味分類毎のキーワード個数を用いて各意味分類についての評価値を決定する。図２（ｂ）は意味分類重要度辞書１５の内容例であり、各意味分類に予め語句の利用目的に応じた重み付け、例えば従来装置と同様の重要度が付与されている様子が示されている。図示の例では、意味分類「１」に重要度”１０”、意味分類「２」に重要度”４”、意味分類「３」に重要度”２”・・・が付与されている。

【００２１】この意味分類評価値決定部１４における評価値の決定手順は図５に示すとおりであり、キーワード意味分類決定部１２から出力された意味分類毎のキーワード個数と、意味分類重要度辞書１５において与えられた当該意味分類の重要度との積で与えられる。なお、これ以外にも、意味分類の重要度とその意味分類に属するキーワードの個数の対数をとったものとの積をその意味分類の評価値とする方法もある。図３（ホ）は意味分類評価値決定部１４の出力例であり、意味分類「１」の評価値が”１２０”、意味分類「３」の評価値が”１０”となる様子が示されている。

【００２２】文書類似度決定部１６では、意味分類カウント辞書１７を参照し、意味分類評価値決定部１４の出力結果に基づいて入力文書に類似する度合い、即ち類似度の高い文書候補を特定する。意味分類カウント辞書１７は、蓄積中の各文書においてどの意味分類に何個のキーワードが含まれているかを個々の文書に対応した文書番号と共に記憶したものである。図２（ｃ）はその内容例を示すものであり、図示の例では、意味分類「１」については文書番号「１２」に４個、文書番号「２４」に５個・・・対応付けられており、意味分類「２」については文書番号「８」に６個・・・が対応付けられている。

【００２３】参照用文書における類似度の尺度として

7

は、例えば、入力文書中に登場する意味分類の評価値と、その意味分類に属するキーワードが文書中で使われている回数の対数をとったものとの積を全ての意味分類について足し合わせたものなどが用いられる。図３（へ）は、このような尺度を用いた場合の文書類似度決定部１６の出力例であり、入力文書に対する類似度の高い順に、文書番号「２４」，「１２」，「１００２」，「６４」・・・に対応する参照用文書が候補になる様子が示されている。なお、文書間の類似度の決定に際しては、全ての意味分類でなく、評価値の大きい意味分類のみを選んで処理することも可能である。

【００２４】次に、上記意味分類カウント辞書１７への意味分類の登録ないし削除手順について説明する。上述の入力文書は、それを次の参照用文書として利用することができる。図６は、上述の入力文書についての処理結果を意味分類カウント辞書１７へ登録する例を示すものである。

【００２５】図６に示されるように、入力文書を形態素解析部１０、キーワードカウント部１１、及びキーワード意味分類決定部１２の順に処理することは上述の実施形態の場合と同様である。即ち入力文書に含まれる意味分類「１」，「３」・・・と各意味分類に属するキーワードの個数とを抽出し、これを新たな文書識別コードである文書番号「１００」と共に意味分類カウント辞書１７に登録する。登録内容を削除する場合は、文書番号「１００」を削除するとともに、各意味分類の数値を減算する。

【００２６】このように、本実施形態の類似文書検索装置では、シソーラス辞書１３を用いて類義のキーワードについての重要度を正しく決定しているので、確度の高い類似文書検索を行うことができ、従来の問題点が解消される。

【００２７】

【発明の効果】上述の説明から明らかなように、本発明によれば、入力文書に正当な重みが付与されるので、文書間の類似度が語句正しく決定され、参照用文書からの類似文書候補の決定精度が高まる、という特有の効果がある。

8

【図面の簡単な説明】

【図１】本発明の一実施形態による類似文書検索装置のブロック構成図。

【図２】（ａ）はシソーラス辞書の内容例、（ｂ）は意味分類重要度辞書の内容例、（ｃ）は意味分類カウント辞書の内容例を示す説明図。

【図３】（イ）は本実施形態による入力文書中の文章例、（ロ）は形態素解析部の出力例、（ハ）はキーワードカウント部の出力例、（ニ）はキーワード意味分類決定部の出力例、（ホ）は意味分類評価値決定部の出力例、（へ）は文書類似度決定部の出力例を示す説明図。

【図４】本実施形態によるキーワード意味分類決定部の処理概要の説明図。

【図５】本実施形態による意味分類評価値決定部の評価値決定過程の説明図。

【図６】本実施形態による意味分類カウント辞書への登録処理の説明図。

【図７】従来の類似文書検索装置のブロック構成図。

【図８】（ａ）は従来装置によるキーワード重要度辞書の内容例、（ｂ）はキーワードカウント辞書の内容例を示す説明図。

【図９】（ト）は従来装置における入力文書中の文章例、（チ）は形態素解析部の出力例、（リ）はキーワードカウント部の出力例、（ヌ）はキーワード評価値決定部の出力例、（ル）は文書類似度決定部の出力例を示す説明図。

【符号の説明】

１０，７１　形態素解析部
１１，７２　キーワードカウント部
１２　キーワード意味分類決定部
１３　シソーラス辞書
１４　意味分類評価値決定部
１５　意味分類重要度辞書
１６，７６　文書類似度決定部
１７　意味分類カウント辞書
７３　キーワード重要度辞書
７４　キーワード評価値決定部
７５　キーワードカウント辞書

【図８】

(a)

73

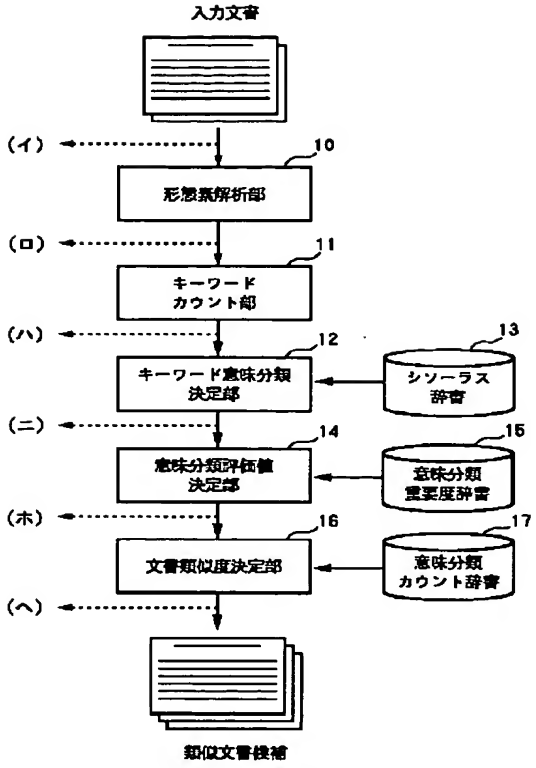
キーワード	意味分類
類似	100
検索	50
文書	10
装置	1
⋮	⋮

(b)

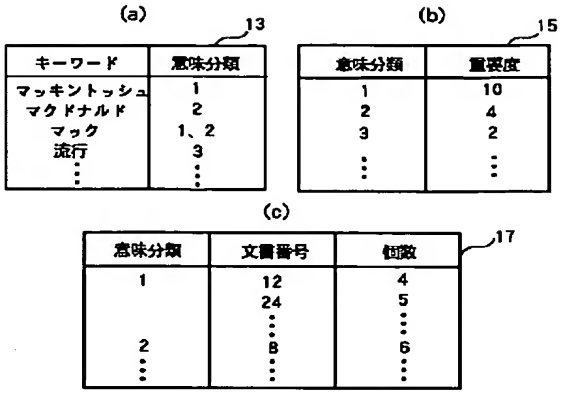
75

キーワード	文書番号	個数
検索	12	4
⋮	24	5
装置	⋮	⋮
⋮	8	6
文書	⋮	⋮
⋮		

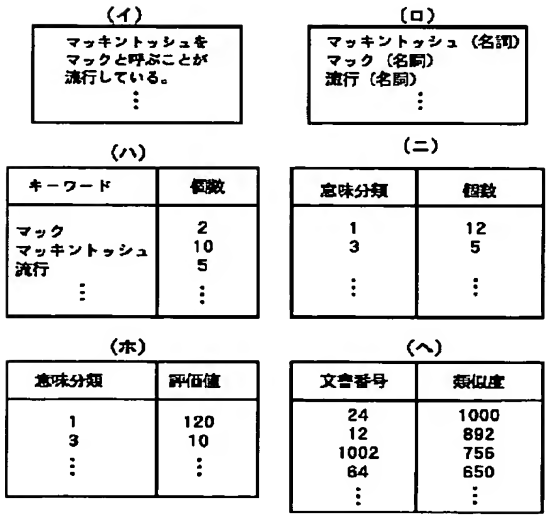
【図１】


入力文書
(イ)
形態素解析部
10
(ロ)
キーワード
カウント部
11
(ハ)
キーワード意味分類
決定部
12
シソーラス
辞書
13
(ニ)
意味分類評価値
決定部
14
意味分類
重要度辞書
15
(ホ)
文書類似度決定部
16
意味分類
カウント辞書
17
(ヘ)
類似文書候補

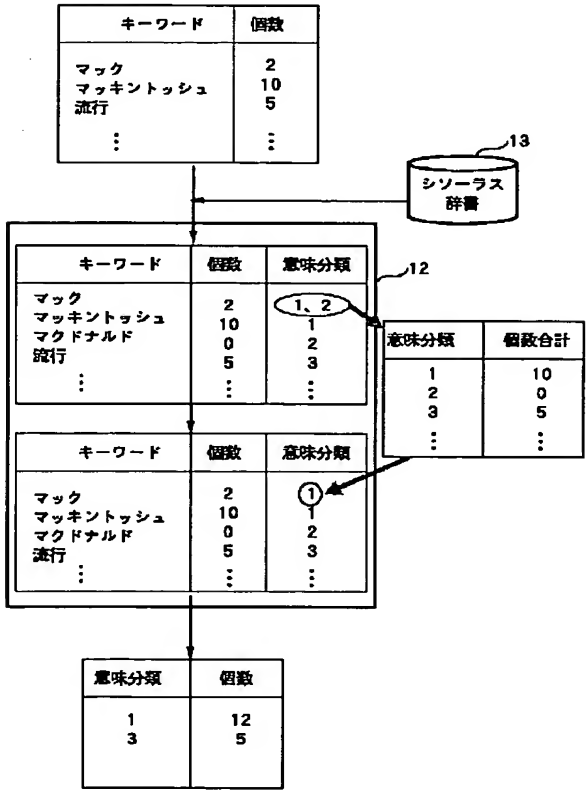
【図２】


(a)
13
キーワード
意味分類
マッキントッシュ
マクドナルド
マック
流行
1
2
1、2
3
(b)
15
意味分類
重要度
1
2
3
10
4
2
(c)
17
意味分類
文書番号
個数
1
2
12
24
8
4
5
6

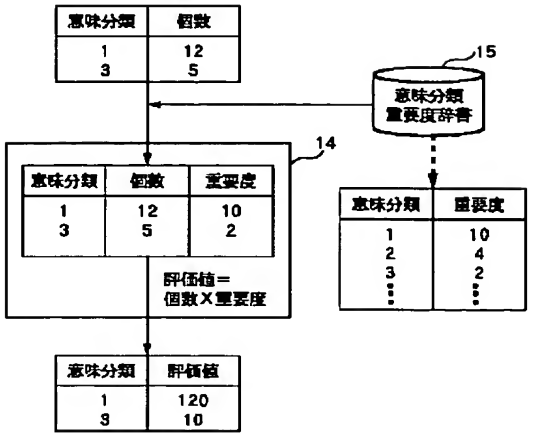
【図３】


(イ)
マッキントッシュを
マックと呼ぶことが
流行している。
(ロ)
マッキントッシュ（名詞）
マック（名詞）
流行（名詞）
(ハ)
キーワード
個数
マック
マッキントッシュ
流行
2
10
5
(ニ)
意味分類
個数
1
3
12
5
(ホ)
意味分類
評価値
1
3
120
10
(ヘ)
文書番号
類似度
24
12
1002
64
1000
892
756
650

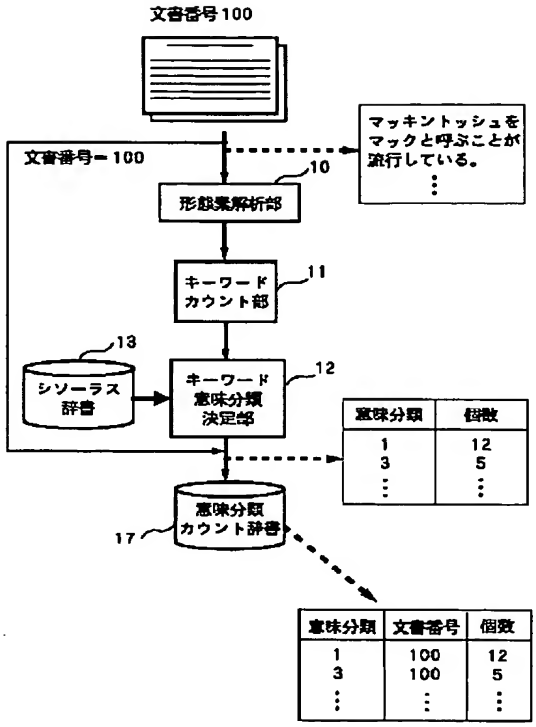
【図４】


キーワード
個数
マック
マッキントッシュ
流行
2
10
5
13
シソーラス
辞書
キーワード
個数
意味分類
マック
マッキントッシュ
マクドナルド
流行
2
10
0
5
1、2
1
2
3
12
意味分類
個数合計
1
2
3
10
0
5
キーワード
個数
意味分類
マック
マッキントッシュ
マクドナルド
流行
2
10
0
5
1
1
2
3
意味分類
個数
1
3
12
5

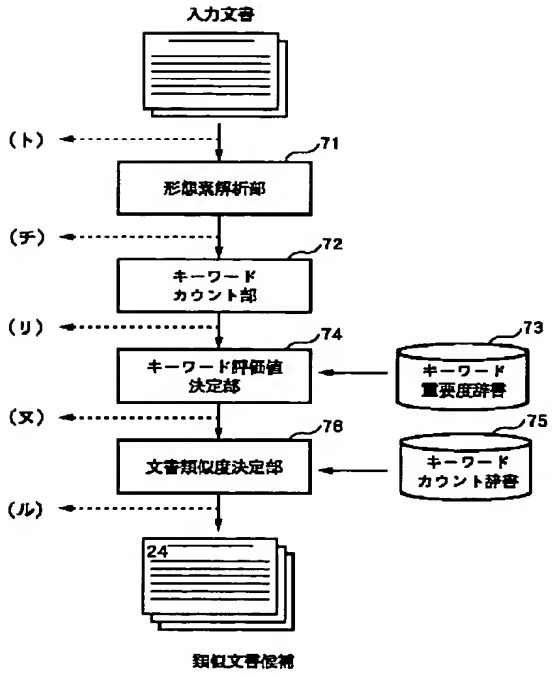
【図５】


意味分類
個数
1
3
12
5
15
意味分類
重要度辞書
14
意味分類
個数
重要度
1
3
12
5
10
2
評価値＝
個数Ｘ重要度
意味分類
重要度
1
2
3
10
4
2
意味分類
評価値
1
3
120
10

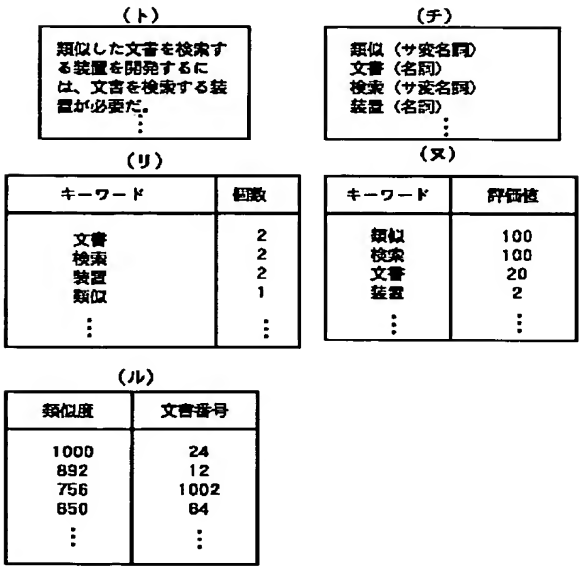
【図６】


文書番号100
文書番号＝100
マッキントッシュをマックと呼ぶことが流行している。
10
形態素解析部
11
キーワードカウント部
13
シソーラス辞書
12
キーワード意味分類決定部
意味分類
個数
1
3
12
5
17
意味分類カウント辞書
意味分類
文書番号
個数
1
3
100
100
12
5

【図７】


入力文書
(ト)
71
形態素解析部
(チ)
72
キーワードカウント部
(リ)
74
キーワード評価値決定部
73
キーワード重要度辞書
(ヌ)
76
文書類似度決定部
75
キーワードカウント辞書
(ル)
24
類似文書候補

【図９】


(ト)
類似した文書を検索する装置を開発するには、文書を検索する装置が必要だ。
(チ)
類似（サ変名詞）
文書（名詞）
検索（サ変名詞）
装置（名詞）
(リ)
キーワード
個数
文書
検索
装置
類似
2
2
2
1
(ヌ)
キーワード
評価値
類似
検索
文書
装置
100
100
20
2
(ル)
類似度
文書番号
1000
892
756
650
24
12
1002
84